# Comparison of the Diversity of Article Keywords and Title Segmentation and Their Knowledge Graph Construction (Postprint)

**Authors:** Li Jihong, Xu Guizhen, Jiang Shan, Hongjiang Wang

**Date:** 2023-10-08T00:00:00+00:00

## Abstract

[ Purpose / Significance ] Based on article keywords and title segmentation, this study analyzes the diversity of both elements and the differences in knowledge graphs constructed from them.

[ Method / Process ] Relevant papers on the research topic of "academic misconduct" were crawled from CNKI (China National Knowledge Infrastructure) covering 2010-2019. Diversity indices were employed to quantitatively analyze the characteristics of article keywords and title segmentation, and CiteSpace software was used to qualitatively compare the architecture of knowledge graphs based on keywords versus title segmentation.

[ Results / Conclusion ] The richness (S), diversity (H' ), and evenness (EH) indices of keywords all differed from those of title segmentation, and weak similarity was observed between the two units, indicating that article keywords and title segmentation constitute distinct units. Although the knowledge graphs constructed based on them showed differences, both could effectively reveal research themes in the "academic misconduct" field from their respective perspectives.

## Full Text

## Preamble

### Diversity of Keywords and Title Segmentation Words and a Comparison of Their Knowledge Mappings

Li Jihong, Xu Guizhen, Jiang Shan, Wang Hongjiang
Institute of Agricultural Economy and Information, Anhui Academy of Agricultural Sciences, Hefei 230031

## Abstract

[**Purpose/Significance**] This study aims to explore the diversity between article keywords and title segmentation words, as well as the differences in knowledge mappings constructed from these two sources. [**Method/Process**] We retrieved academic papers on the theme of "academic misconduct" from CNKI (CSSCI source journals) spanning 2010–2019. Using diversity indices, we quantitatively analyzed the characteristics of keywords and title segmentation words, and qualitatively compared their knowledge mapping architectures through CiteSpace visualization software. [**Result/Conclusion**] The results show that the richness index (S), Shannon-Wiener diversity index (H ), and Pielou evenness index (EH) of keywords all differed from those of title segmentation words, with weak similarity between the two units, indicating that keywords and title segmentation words represent distinct analytical units. Although the resulting knowledge mappings showed differences, both approaches effectively revealed research themes in the "academic misconduct" field from their respective perspectives.

**Keywords:** academic misconduct; keyword; title; Chinese word segmentation; diversity; knowledge mapping

---

## Introduction

Knowledge mapping integrates theories and methods from applied mathematics, graph theory, information science, and other disciplines with bibliometric approaches such as citation analysis and co-occurrence analysis. Through visual mapping, it reveals the core structure, developmental history, frontier areas, and overall knowledge architecture of a discipline, achieving multidisciplinary integration. This approach can display complex knowledge domains through data mining, information processing, knowledge measurement, and graphical representation, uncovering dynamic developmental patterns and providing valuable references for disciplinary research [1]. Current knowledge mapping tools include CiteSpace, HistCite, Sci2 Tools, Leydesdorff, Ucinet, Pajek, VOSviewer, and others [2]. Among these, CiteSpace is the most widely used and functionally powerful information visualization software, enabling analysis of co-citation networks, co-occurrence networks, or collaboration networks through selectable node types to generate visualized, sequential knowledge maps [3-5].

In published CiteSpace-related papers, a significant proportion involves co-occurrence analysis of keywords. Keywords are words or terms selected from reports and papers for literature indexing purposes to represent the main content of the full text [6]. As the "abstract of the abstract," they highly summarize the paper' s theme and concentrate the core essence of its content. Co-occurrence analysis of keywords can explore research hotspots, trends, and knowledge structures in a field. For data sources without keywords (such as paper titles, fund projects, online public opinion, or film reviews), Chinese

word segmentation is primarily employed [7-9]. Segmentation is the process of reorganizing a continuous character sequence into word sequences according to specific rules. Paper titles are brief statements indicating article content and constitute the most important part of a paper. Segmenting titles yields words that cover article content and main ideas.

Both paper keywords and title segmentation words can reveal thematic content. For the same article, what is the relationship between keywords and title segmentation words, and what differences exist in the knowledge mappings constructed from them? This study uses CNKI (CSSCI source journals) as the data source, retrieving research literature on "academic misconduct." Using diversity indices for quantitative analysis of keywords and title segmentation words, and employing CiteSpace software for qualitative comparison of mapping architectures based on these two word sources, this research provides both theoretical foundations and practical significance for related studies.

# 1. Methods

## 1.1 Data Sources

We retrieved academic papers on the theme of "academic misconduct" from CNKI (CSSCI source journals). The search query was: "Subject = academic misconduct OR academic fraud OR academic ethics OR research integrity," with publication years from 2010 to 2019. The retrieval date was October 24, 2019. After removing notices, calls for papers, and declarations, we obtained 757 academic papers. We then crawled and compiled bibliographic information, including paper titles, authors, and keywords.

### 1.2.1 Title Segmentation

Title segmentation should follow these principles: (1) Select words that clearly express thematic concepts; (2) Remove meaningless words; (3) Merge vocabulary with identical meanings but different expressions. For example, after standardization, terms such as academic misconduct and academic corruption were unified as "academic misconduct behavior," while AMLC and academic misconduct detection systems were unified as "academic misconduct literature detection system."

Based on these principles, we used the ROSTCM6 software developed by Wuhan University to segment titles. Directly obtained keywords had relatively coarse granularity and unsatisfactory clustering effects, so we supplemented user dictionaries through manual annotation to meet research requirements. A data cleaning tool was then used to clean and merge segmented words.

### 1.2.2 Diversity Analysis of Keywords and Title Segmentation Words

Diversity is a common concept in ecology. Commonly used diversity indices include the richness index (S), Shannon-Wiener diversity index (H ), and Pielou

evenness index (EH), which are used to assess the diversity and complexity of communities or ecosystems [10-12]. This study borrows these three indices to describe the diversity of keywords and title segmentation words.

Formulas (1) and (2) are used for calculation, where S represents the number of keywords or title segmentation words; Pi represents the relative abundance of word i, indicating the proportion of a particular word' s count in the total word count, i.e., Pi = ni/N, where ni is the count of word i, and N is the total count of all keywords or title segmentation words.

To examine similarity between the two units, we employed the Sørensen index (CS) and Jaccard index (CJ) for analysis. These ecological indices reflect species composition similarity between communities [13] and are here used to reflect word similarity between units.

In these formulas, c represents the number of shared words between keywords and title segmentation words; a and b represent the word counts of keywords and title segmentation words, respectively.

#### 1.2.3 Format Conversion for Title Segmentation Words

As is well known, CiteSpace software can only analyze literature from specific databases and cannot directly process data from other sources. We used format conversion software to transform data from non-specific databases into formats recognizable by CiteSpace for analysis.

#### 1.2.4 Co-occurrence Analysis of Keywords and Title Segmentation Words

Co-occurrence analysis examines keyword or title segmentation word sets in a dataset. Through visualization analysis, it can identify disciplinary structures and research hotspots in a field. We used "academic misconduct" research literature keywords and title segmentation word sets as analysis units, employing CiteSpace to generate thematic clustering maps to compare knowledge architectures in this domain.

Visualization parameters were set as follows: timespan 2010–2019, time slicing = 1 year, node types = keyword, node strength defaulting to Cosine and Within Slices options, selection criteria using Top N per slice = 50, and pruning function defaulting to no pruning. This generated co-occurrence knowledge maps for both keywords and title segmentation words.

## 2. Results

### 2.1 Word Composition of Keywords and Title Segmentation Words

We statistically analyzed keywords and title segmentation words from "academic misconduct" literature, obtaining 3,131 keywords and 3,094 title segmentation

words. After consolidating identical items from each unit, we ultimately identified 1,541 unique keyword terms and 1,432 title segmentation word terms.

The frequency distribution and word counts at each frequency level are shown in Table 1 . The keyword with highest frequency (344 occurrences) was "academic misconduct behavior," followed by "research integrity" (95) and "academic ethics" (77). The minimum frequency was 1. As frequency decreased, the number of keywords at that frequency increased. For example, high-frequency words like academic misconduct behavior, research integrity, academic ethics, and graduate student had fewer words at their frequency levels (1 each), while keywords with frequencies of 3 or below were numerous. There were 147 keywords with frequency = 2 and 1,254 with frequency = 1, accounting for 40% of the total. The most frequent title segmentation word was also "academic misconduct behavior" (268 occurrences), followed by "graduate student" (107) and "university" (84). The minimum frequency was also 1, with 1,094 words occurring only once, representing 35.36% of all title segmentation words.

For "academic misconduct" research, keywords included 37 more words than title segmentation words, with 109 more total terms (7.61%), indicating that paper-provided keywords are richer than title segmentation words. However, high-frequency words were relatively consistent between both approaches, including academic misconduct behavior, graduate student, university, research integrity, and scientific journals, suggesting that core terms remain constant regardless of method.

### 2.2 Diversity of Keywords and Title Segmentation Words

Based on word frequencies and counts, we analyzed the diversity of these two units. Since CiteSpace visualization requires a frequency threshold $ $2, we analyzed diversity for all words and for words with frequency $ $2 separately.

We primarily used richness index (S), Shannon-Wiener diversity index (H ), evenness index (EH), Sørensen index (CS), and Jaccard index (CJ) to analyze diversity from both intra-unit and inter-unit perspectives, exploring differences in quantity, abundance, and distribution of words expressing the same theme from different sources, as well as their similarity. Richness, Shannon-Wiener, and evenness indices are $\alpha$-diversity indices measuring intra-unit structural diversity. Richness index (S) reflects word quantity; Shannon-Wiener index (H ) reflects intra-unit diversity based on word counts, with higher H values indicating greater information content and diversity; evenness index (EH) reflects distribution uniformity, with higher EH values indicating more equal word counts. Sørensen and Jaccard indices are $\beta$-diversity indices measuring inter-unit similarity, with higher values indicating greater similarity and consistency.

Diversity measures are shown in Table 2 . For all keywords, richness index was 1,541, Shannon-Wiener index was 6.25, and evenness index was 0.85. For all title segmentation words, richness was 1,432, Shannon-Wiener index was 6.26, and evenness index was 0.96. Keywords showed greater richness than title segmen-

tation words, with similar Shannon-Wiener indices, but lower evenness, as title segmentation words had more uniform distribution. Since words with frequency = 1 were numerous, richness decreased substantially for words with frequency $ $2: keyword richness dropped 81.38% (to 287), while title segmentation word richness decreased 76.40% (to 338). For these higher-frequency words, keywords showed lower richness (287) and diversity (4.54) than title segmentation words (338 and 4.86, respectively), but similar evenness. CS and CJ measured similarity between units. For all frequencies, similarity was low (CS = 0.39, CJ = 0.24). For frequency $ $2 words, similarity was slightly higher (CS increased 17.95% to 0.46, CJ increased 25% to 0.30), but still indicated weak correlation.

## 3. Comparison of Knowledge Mappings

### 3.1 Graph Parameter Comparison

Based on keywords and title segmentation words from 2010–2019 publications and using uniform parameter settings, co-occurrence networks were generated as shown in Figure 1 [Figure 1: see original paper]. The software selected the top 50 most cited words per year, with each word appearing at least twice. In Figure 1, numbers represent cluster IDs, and node size indicates word frequency.

Graph parameters displayed in the upper-left corner include software information, runtime, data location, timespan, selection criteria, network generation, pruning method, modularity, and mean silhouette value. Modularity and mean silhouette reflect overall graph characteristics. Table 3 presents parameters for both mappings, including N (nodes), E (links), density, modularity, and silhouette. The keyword-based mapping had 115 nodes and 327 links with density 0.0499. The title segmentation mapping had 131 nodes and 420 links, representing increases of 13.91% and 28.44% respectively, with similar density (0.0493). Modularity and silhouette are crucial parameters. The keyword mapping showed modularity 0.4352 and silhouette 0.3667. The title segmentation mapping showed modularity 0.4115 (5.7% lower) and silhouette 0.4935 (34.58% higher), indicating higher cluster homogeneity. Both modularity values exceeded 0.3, the threshold for significant clustering structure, confirming that both approaches produce valid cluster structures.

### 3.2 Word Comparison in Mappings

Word frequency and betweenness centrality indicate term importance in the mapping. Betweenness centrality quantifies node importance in networks [2]; higher values indicate greater significance. In the keyword mapping, 9 words had frequency $ $30, totaling 782 occurrences. "Academic misconduct behavior" ranked highest (340 occurrences, betweenness centrality 0.38), followed by "research integrity" (92, 0.23), "academic ethics" (76, 0.20), and "graduate student" (67, 0.32). While frequency and centrality rankings didn' t perfectly correspond, they were generally consistent. In the title segmentation mapping, 7 words had frequency $ $30, totaling 639 occurrences—fewer than the keyword

mapping. However, main terms were similar, including academic misconduct behavior, research integrity, academic ethics, graduate student, scientific journals, and university. The highest-frequency term was again academic misconduct behavior (262, betweenness centrality 0.39).

### 3.3 Cluster Comparison

Cluster analysis is an exploratory data mining method for identifying and analyzing significant term categories and contexts in specific research fields. It uses algorithms to structure data into clusters, revealing thematic distributions and organizational structures [14].

Table 4 shows words with frequency $ $30 in both mappings, including term frequency and betweenness centrality. The keyword mapping formed 6 clusters: #0 academic ethics, #1 executive editor, #2 academic journals, #3 dissertations, #4 research management, and #5 anomie. The title segmentation mapping also formed 6 clusters: #0 academic misconduct behavior, #1 research integrity, #2 practice, #3 academic ethics anomie, #4 retraction, and #5 MOOCs. Table 5 presents cluster information, including nodes per cluster and silhouette values. In the keyword mapping, all clusters had high silhouette values. In the title segmentation mapping, all clusters except #1 (silhouette 0.375) showed high homogeneity. Combined with Figure 1 parameters, both mappings demonstrate ideal clustering.

Based on semantic structure and thematic relevance, clusters can be integrated. The keyword mapping integrates into three knowledge domains: (1) academic misconduct behavior and anomie (#0, #5), covering specific manifestations; (2) executive editor and academic journals (#1, #2), reflecting the journal focus; (3) dissertations and research management (#3, #4), showing universities as another key research area. The title segmentation mapping also integrates into three domains: (1) academic misconduct behavior, research integrity, and academic ethics anomie (#0, #1, #3), focusing on core issues; (2) practice and MOOCs (#2, #5), covering prevention through literacy education and system construction; (3) retraction (#4), examining consequences and impacts of retracted papers.

Both mappings share similarities and differences in cluster structure, related to word sources. Keywords include both core and non-core thematic terms—those limiting/modifying core concepts, specific research content, new methods, and temporal/spatial factors—together summarizing themes and core content. Title segmentation words, derived from titles that distill and condense essential content, mostly represent core thematic terms with fewer non-core supplementary terms, potentially lacking some modifiers. Consequently, both mappings clearly and objectively display research themes in academic misconduct, but with different emphases: keyword-based mappings focus more on research problems, while title segmentation mappings emphasize research methods and approaches.

## 4. Discussion and Conclusion

**(1) Expanded Application of CiteSpace Software.** How to quickly clarify research architectures from massive literature, identify the most critical effective information, and understand past, present, and future trends is a major challenge in scientific research. Knowledge mapping provides a valuable scientific approach to this problem. The information visualization software CiteSpace is a powerful tool whose mappings offer "one map showing history at a glance" characteristics [2], gaining widespread application since its inception. Currently, CiteSpace can only analyze literature from specific databases (WoS, Scopus, ADS, arXiv, CNKI, CSSCI, NSF, CSCD, Derwent patents, etc.) and cannot directly process data from other sources. We used format conversion software to transform non-specified database data into CiteSpace-compatible formats. Results demonstrate this method is scientifically effective, expanding CiteSpace' s data sources and providing references for visualizing data from non-specified databases.

**(2) Diversity of Keywords and Title Segmentation Words.** Language, as a tool for logical thinking and reasoning, has words as its basic elements [16]. Using keywords and title segmentation words from academic literature as conceptual evolution foundations, we employed richness index (S), Shannon-Wiener index (H ), evenness index (EH), Sørensen index (CS), and Jaccard index (CJ) to compare diversity between these word types.

For academic misconduct research, keywords included 109 more terms than title segmentation words, though high-frequency terms remained consistent. This indicates that core terms are identical regardless of method. For all words, keywords showed greater richness than title segmentation words, with similar diversity indices but lower evenness due to less uniform distribution. Excluding numerous single-occurrence words, richness and diversity for frequency $ $2 words decreased substantially for both approaches, with keywords showing lower richness and diversity than title segmentation words but similar evenness. CS and CJ indicated weak similarity between units for both all-frequency and high-frequency words, confirming they are substantially different units—a finding that supports subsequent knowledge mapping.

**(3) Co-occurrence Networks of Keywords and Title Segmentation Words.** Keywords facilitate literature indexing and retrieval, representing full-text thematic content. CiteSpace automatically extracts keywords from conventional databases, including both core and non-core thematic terms. This study additionally obtained words through title segmentation, yielding mostly core thematic terms with fewer non-core terms. Both approaches contain core and non-core terms but differ in quantity and content, resulting in mappings with both similarities and differences. Both clearly and objectively display academic misconduct research themes, but with distinct emphases: keyword-based mappings focus on research problems, while title segmentation mappings emphasize research methods and approaches.

## References

[1] Yang Siluo, Han Ruizhen. Analysis of methods and tools for knowledge mapping abroad [J]. Library and Information Service, 2012(6): 101-109.
[2] Li Jie, Chen Chaomei. Citespace: Technology Text Mining and Visualization [M]. Beijing: Capital University of Economics and Business Press, 2016.
[3] Chen C M, Ibekwe-SanJuan F, Hou J H. The structure and dynamics of co-citation clusters: a multiple-perspective co-citation analysis[J]. Journal of the American Society for Information Science and Technology, 2010, 61(7): 1386-1409.
[4] Hou Jianhua, Yang Xiucai, Zhou Lijuan. Analysis of frontier topics and evolution trends in international library and information science research [J]. Library and Information Service, 2016, 60(13): 82-90.
[5] Si Hongyun, Shi Jiangang, Chen Jindao, et al. Examining domestic sustainable development research from the perspective of *China Population, Resources and Environment*: thematic threads, knowledge evolution, and emerging hotspots [J]. China Population, Resources and Environment, 2019, 29(7): 166-176.
[6] Presentation of scientific and technical reports, dissertations and scientific papers: GB 7713—87[S]. Beijing: National Documentation Standardization Technical Committee, 1987.
[7] Chen Bikun, Zhan Changjing. Visual analysis of the "library, information and archives management" disciplinary structure from the perspective of national fund projects [J]. Journal of Intelligence, 2017, 36(7): 105-110.
[8] Ren Zhongjie, Zhang Peng, Li Sicheng, et al. Analysis of emotional situation evolution in emergencies based on Weibo data mining: a case study of the Tianjin 8·12 accident [J]. Journal of Intelligence, 2019, 38(2): 140-148.
[9] Shan Xiaohong, Pang Shihong, Liu Xiaoyan, et al. Analysis of network public opinion evolution paths based on event logic graphs: a case study of medical public opinion [J]. Information Studies: Theory & Application, 2019, 42(9): 99-103, 85.
[10] Lai Baochun, Dai Ruiqing, Wu Zhenqiang, et al. Comparative study on bacterial community diversity in rhizosphere soil of healthy and fusarium wilt-infected pepper plants [J]. Fujian Journal of Agricultural Sciences, 2019, 34(9): 1073-1080.
[11] Liu Zheng, Li Ying, Zhu Pei, et al. Diversity and flora of wetland vascular plants in Changxing County, Zhejiang Province [J]. Journal of Zhejiang A&F University, 2020, 37(3): 465-473.
[12] Zhou Yuzhen, Wu Pengfei. Diversity and spatiotemporal distribution of forest soil microarthropod communities on the eastern slope of Gongga Mountain [J]. Chinese Journal of Ecology, 2020, 39(2): 569-576.
[13] Ji Changpin, Wang Hua. Distribution characteristics of wetland plant communities in Poyang Lake and their response to soil environmental factors [J]. Ecology and Environmental Sciences, 2018, 27(8): 1417-1424.
[14] Olawumi T O, Chan D W M. A scientometric review of global research on sustainability and sustainable development[J]. Journal of cleaner production,

2018, 183: 231-250.
[15] Academic publishing specification—Rules for keyword compilation: CY/T 173—2019[S]. Beijing: National Press and Publication Administration, 2019.
[16] Huang Chen, Zhao Xing, Bian Yangyi, et al. Analysis of keyword analysis method for measuring academic contribution [J]. Journal of Library Science in China, 2019, 45(6): 84-99.

## Author Contributions

Li Jihong: Conceptualized the study, analyzed data, and wrote the manuscript;
Jiang Shan: Collected data;
Xu Guizhen: Provided revision suggestions;
Wang Hongjiang: Provided revision suggestions.

*Note: Figure translations are in progress. See original paper for figures.*